MSC önálló laboratórium 2. beszámoló jegyzet

# Research and Technologies

## The role of deep learning in spatial transcriptomics

### Unsupervised spatially embedded deep representation of spatial transcriptomics literature review

Spatial transcriptomics is a method that measures gene activities/expressions directly from tissue samples, keeping track of where each gene is active. New methods are measuring genes while keeping cells’ spatial location inact. This helps in:

* Understanding how different cells interact in their natural environment
* Studying complex diseases like cancer or brain disorders better

**Model structure**

Unsupervised spatially embedded deep representation (SEDR) method is for learning a low-dimensional latent representation of gene expression embedded with spatial information. The SEDR model consists of two main components:

* Masked self-supervised deep autoencoder network → learns latent representation
* Variational graph convolutional autoencoder (VGAE) network → embeds spatial information

The low-dimensional embedding produced by SEDR can be used for:

* Downstream visualization
* Spot clustering
* Trajectory inference
* Batch effect correction

The reconstructed feature matrix can be used to impute the raw gene expression with dropouts.

**Data masking**

Generating a masked gene expression matrix as input → fed into the deep autoencoder + VGAE.

1. The matrix has rows as spots (locations on tissue) and columns as different genes. Each element tells you the activity of the gene at a specific spot.
2. Randomly selects some of these spots and temporarily hides their gene expression data by replacing their values with placeholders.
3. The model reconstructs the original gene expressions based on only the visible spots and their spatial relationships.
4. The quality is measured by mean squared error (MSE).

**Graph construction for spatial transcriptomics data**

This method builds a spatial graph to capture how spots relate to each other based on their location.

1. Calculate Euclidean distances between spots.
2. K-nearest neighbors for each spot to create adjacency matrix (1 if neighbors, 0 if not).
3. The matrix is stored as a sparse matrix (remember just the connected pairs) to save computing resources.

**Latent representation learning**

Main learning step, involving deep neural networks.

1. Deep autoencoder (Gene Expression)
   * Compresses the masked gene expression into smaller set of features → low-dimensional embedding.
   * Compressed data is then decoded.
   * Difference between reconstructed and original is minimized.
2. VGAE (Spatial information)
   * Takes the adjacency matric and gene embedding to create a spatially informed embedding
   * First, calculates average and variation using a Graph Convolutional Network (GCN).
   * Second, creates the spatial embedding from these parameters using randomness (reparameterization) to ensure diversity in spatial representation.
   * Reconstructs the adjacency matrix, ensuring the spatial structure is captured.
3. Combining gene and spatial information
   * The simplified gene embedding and spatial embedding are merged to form a complete latent representation.

**Batch efect correction for spatial transcriptomics**

Spatial datasets often come from different labs or conditions, causing unwanted technical differences called batch effects.

1. Multiple datasets are combined into a larger block-diagonal adjacency matrix, with no direct connection between spots from different datasets.
2. This is further processed using an advanced clustering method called Deep Embedded Clustering (DEC).
   * Spots are grouped into clusters automatically and iteratively refined.
   * DEC improves clusters by using a soft-assignment method based on how similar spots are in the latent representation.
   * DEC further uses a "Student’s t-distribution" to softly group spots into clusters, optimizing these groups by minimizing the differences (KL divergence) between predicted and ideal cluster assignments.

**Clustering**

After getting the simplified latent representation to group similar spots, revealing hidden patterns in tissues.

From various clustering methods, mclust was used due to high performance. This clearly organizes the spots based on the learned latent representation.

**Application and validation on real-world data**

SEDR was tested with real spatial datasets:

* Human brain data (clearly identified brain layers).
* Human cancer data (ovarian, breast) and lymph node data (denoising and filling missing data).
* High-resolution mouse tissue data (Stereo-seq and Slide-seq) (clearly separated fine tissue structures, proved efficient and scalable).

### Overview of clustering and denoising problems

**Clustering problems in spatial transcriptomics**

1. Ignoring spatial context (spatially naïve clustering)
   * Cells physically close often share similar biological characteristics → traditional clustering methods like K-means, ignores spatial location data → loss of important information.
   * Solutions: GNN-based clustering → encodes spatial context into representation
     + Spatial Graph Convolutional Networks (GCNs)
     + Spatial Variational Autoencoders (VAEs)
     + GraphST, SEDR, and STAGATE methods → explicitly incorporate spatial context to cluster accurately
2. Difficulty identifying rare cell types (imbalanced clusters)
   * Rare cells often have subtle expression signals and are challenging to cluster.
   * Solutions:
     + Graph Attention Networks (GATs) → weigh neighboring cells differently, potentially highlighting subtle differences
     + Oversampling/Undersampling methods
3. Selecting the number of clusters (hyperparameter tuning)
   * Solutions:
     + Heuristic methods (Silhouette scores, Elbow method, biological validation via marker genes)
     + VAE or self-supervised methods can identify optimal latent dimensionalities, indirectly influencing cluster granularity

**Low-expression filtering (Technical noise)**

Many genes have very low expression, sometimes due to technical errors, not biological reality, which introduce noise. This can be solved:

* Threshold-based filtering → gene expressed in very few cells/spots are removed
* Statistical methods → modeling the expected distribution of gene expression and filtering outliers or unreliable low signals
* Deep learning-based denoising (AE/VAE) → these learn robust latent representations that can reconstruct a less noisy expression profile, improving data quality significantly

**Batch effect correction (Technical noise)**

Spatial transcriptomics data generated from different experimental runs, conditions, or equipment can have systematic differences. Solutions:

* Statistical methods (Harmony, ComBat) → adjust the data distribution to align batches
* Deep learning methods → graph-based autoencoders, combine spatial and gene expression data across batches into a consistent, biologically meaningful embedding
* Methods like SEDR or STAGATE → specifically consider batch effects and attempt to correct them during the embedding phase

### Exploration of existing GNN-based methods

* **SEDR**
  + Deep autoencoder
  + VAE
* **GraphST**
  + Graph self-supervised contrastive learning framework to integrate gene expression and spatial data.
  + Learns spot representations that enhance spatial clustering, data integration, and cell-type deconvolution tasks.
* **STAGATE** 
  + Like SEDR but uses graph attention auto-encoder framework to adaptively learn the similarity between neighboring spots.
* **STdGCN (Spatial Transcriptomic Deconvolution using GCN)**
  + Integrates single-cell RNA sequencing (scRNA-seq) data with spatial transcriptomics for deconvolution.​
  + Utilizes GCNs to model spatial relationships.
* **DenoiseST**
  + Employs a dual-channel learning strategy with GCNs.
  + Enhances discriminative information learning from global data distributions.
  + Adaptively fits different gene distributions to clustered domains.

## Selection of publicly available spatial transcriptomics datasets

### 10x Genomics Visium datasets

* High-resolution spatial gene expression data.​
* Associated histological images.​
* Metadata detailing experimental conditions.

### Spatial research published datasets

* Gene expression matrices with spatial coordinates.​
* Corresponding histological images.​
* Detailed experimental metadata.​

### STOmicsDB

* Spatial transcriptomics datasets from various species and tissues.​
* Standardized data formats for ease of use.​
* Metadata including experimental protocols and conditions

### SpatialDB

* Curated spatial transcriptomic datasets from various studies.​
* Annotations and metadata for each dataset.​
* Tools for data visualization and analysis.

### SOAR (Spatial transcriptOmics Analysis Resource)

* A comprehensive database comprising 441 spatial transcriptomics datasets from 13 species across 42 tissue types.
* It provides a platform for evaluating spatial variability of genes, assessing cell-cell interactions, and visualizing spatial gene expression.

## Learning about Python libraries

### PyTorch Geometric

Graph Neural Networks

### Scanpy

Analysis of single-cell data

### Squidpy installation and configuration

Spatial transcriptomics extension to Scanpy

# Data preparation

## Data loading and preprocessing

### Handling gene expression matrices and spatial coordinates

### Basic noise reduction (low-expression filtering, batch effect correction)

## Graph representation construction

### Transforming transcriptomics and cells into graph structures

### Evaluating various graph structures (k-NN graph, Delaunay triangulation)

### [1] Xu, H., Fu, H., Long, Y. *et al.* Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Med* 16, 12 (2024). <https://doi.org/10.1186/s13073-024-01283-x>

### [2] Dong, K., Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 13, 1739 (2022). <https://doi.org/10.1038/s41467-022-29439-6>

[3] Cui, Yaxuan & Wang, Ruheng & Zeng, Xin & Cui, Yang & Zhu, Zheyong & Nakai, Kenta & Ye, Xiucai & Sakurai, Tetsuya & Wei, Leyi. (2024). DenoiseST: A dual-channel unsupervised deep learning-based denoising method to identify spatial domains and functionally variable genes in spatial transcriptomics. <https://doi.org/10.21203/rs.3.rs-4470472/v1>

[4] Li, Y., Luo, Y. STdGCN: spatial transcriptomic cell-type deconvolution using graph convolutional networks. *Genome Biol* 25, 206 (2024). <https://doi.org/10.1186/s13059-024-03353-0>

[5]